Fairness, Accountability, and Transparency

Machine Learning: Jordan Boyd-Graber
University of Maryland
BIASED REPRESENTATIONS

Slides/ideas adapted from Adam Tauman Kalai and Moritz Hardt
Our data reflect our world …

- Word representations learned from massive amounts of data
- Reflect prejudices and messiness of our world
- But learned representations used for many tasks
  - Detecting “bad” behavior online
  - Matching resumes to jobs
  - Recommendations
The embedding captures gender stereotypes and sexism. (related [Schmidt ‘15])
Easier to debias an embedding than to debias a human

DEFINITIONAL

(state related [Schmidt ‘15])
Consistency of embedding stereotype

GloVe trained on web crawl

word2vec trained on Google news

Doesn’t matter source or algorithm

Each dot is an occupation; Spearman = 0.8
Bias encoded in some dimensions
Analogies

he:$x$:she:$y$

$$\min \cos(he - sh, x - y) \text{s.t.} ||x - y||_2 < \delta$$ (1)

29/150 analogies rated as gender stereotypic by majority of crowdworkers
Bias Where it Shouldn’t Be
Debiasing

218 gender-definition words

Linear SVM

he, she, he, king, queen, programmer, smart, blue, pink, cute, homemaker
Debiasing

Diagram showing vectors $B$ and $B_{⊥}$ with words such as "he", "she", "king", "queen", "programmer", "smart", "blue", "pink", "cute", "homemaker".
Debiasing
Debiasing

# stereotypic analogies

# appropriate analogies

# analogies generated
Debiasing

Original embedding

softball

pitcher

receptionist

maestro

footballer

football

Debiased embedding

softball

pitcher

major leaguer

midfielder

footballer

football
Data are biased …

- Our data (societies) are biased
- Can we make algorithms better than the data?
- Can we define fairness for tasks like sentencing, loan approval, etc.
Defining Fairness

What does non-discriminatory mean?

Target \( y \), predictor \( \hat{y} \) from features \( x \) and protected attribute \( a \).

- Don’t want to remove \( a \)
- Don’t want parity \( p(\hat{y} | A = a) = p(\hat{y} | A = a') \)
Defining Fairness

What does non-discriminatory mean?

Target $y$, predictor $\hat{y}$ from features $x$ and protected attribute $a$.

- Don’t want to remove $a$ (correlations, accuracy disparity)
- Don’t want parity $(p(\hat{y} | A = a) = p(\hat{y} | A = a'))$
Defining Fairness

What does non-discriminatory mean?

Target $y$, predictor $\hat{y}$ from features $x$ and protected attribute $a$.

- Don’t want to remove $a$ (correlations, accuracy disparity)
- Don’t want parity $p(\hat{y} | A = a) = p(\hat{y} | A = a')$ (doesn’t allow perfect prediction)
  Also, can have accuracy disparity: give loans to qualified $A = 0$ and random $A = 1$
Defining Fairness

What does non-discriminatory mean?

Target $y$, predictor $\hat{y}$ from features $x$ and protected attribute $a$.

- Don’t want to remove $a$ (correlations, accuracy disparity)
- Don’t want parity ($p(\hat{y} | A = a) = p(\hat{y} | A = a')$) (doesn’t allow perfect prediction)
- Equalized odds:

$$p(\hat{y} | Y = y, A = a) = P(\hat{y} | Y = y, A = a')$$  \hspace{1cm} (2)

- Perfect predictor always satisfies
- Protects against accuracy disparity